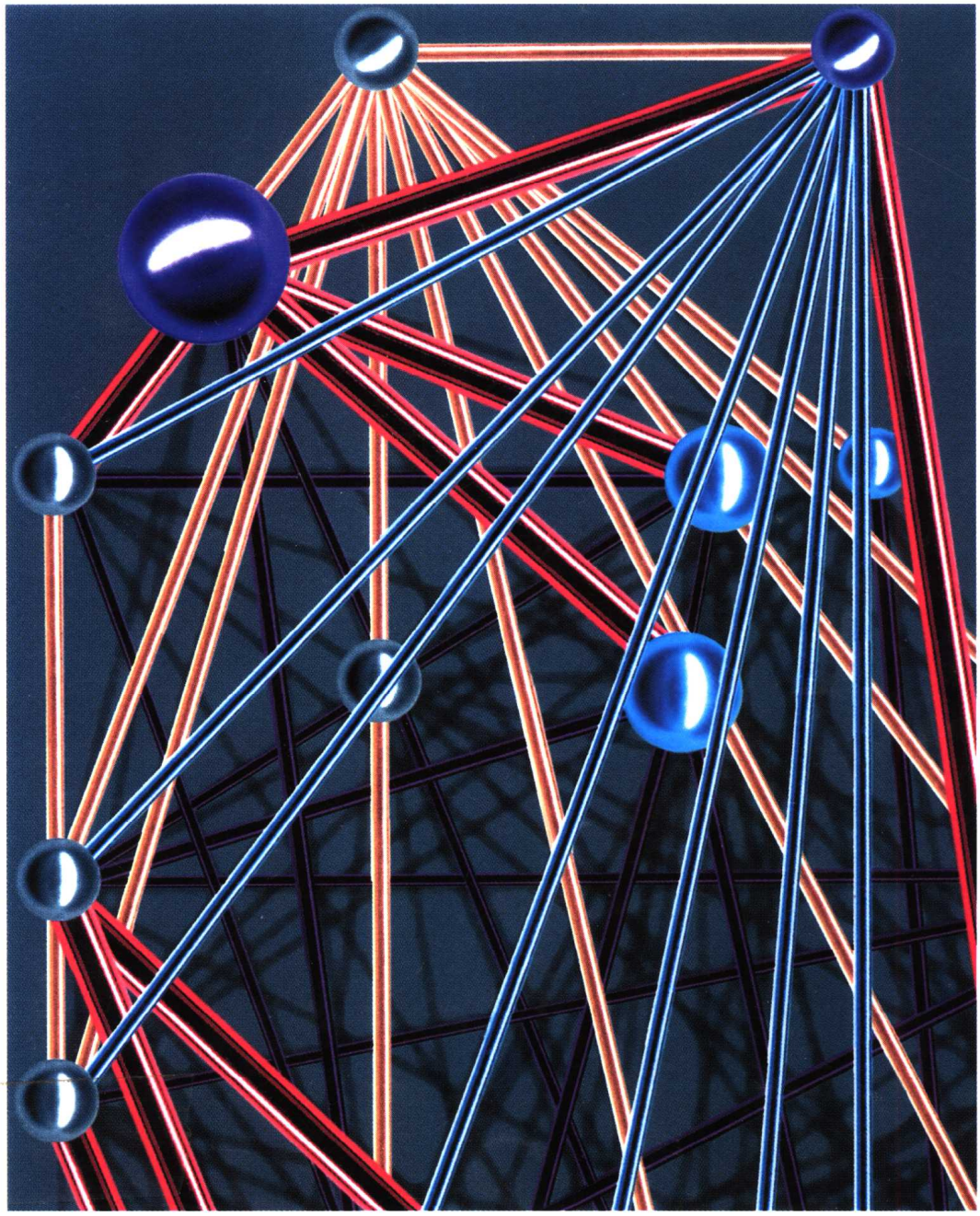
生物信息学 中文版

钟 扬 王 莉 张 亮 主译 李亦学 钱晓茵 张晓宁 校



[美] David W. Mount

高等教育出版社

生物信息学

Bioinformatics: Sequence and Genome Analysis

[美] David W. Mount 著

钟 扬 王 莉 张 亮 主译

李亦学 钱晓茵 张晓宁 校

参加翻译人员：

钟 扬 王 莉 张 亮 沈 玮 朱 彬

田春杰 张晓艳 张文娟 耿宇鹏 戎 俊

雷一东 高 虹 王利民

高等教育出版社

图字:01 - 2002 - 1560 号

Bioinformatics: Sequence and Genome Analysis

David W. Mount

图书在版编目(CIP)数据

生物信息学/[美]芒特(Mount D. W.)著;钟扬等译.

—北京:高等教育出版社,2003.9

ISBN 7-04-012187-5

Ⅰ.生... Ⅱ.①芒...②钟... Ⅲ.生物信息论 Ⅳ.Q811.4

中国版本图书馆 CIP 数据核字(2003)第 058271 号

策划编辑 邹学英 **责任编辑** 吕庆娟 **封面设计** 张 楠 **责任绘图** 朱 静

版式设计 陆瑞红 **责任校对** 康晓燕 **责任印制** 陈伟光

出版发行 高等教育出版社

社 址 北京市西城区德外大街 4 号

邮政编码 100011

总 机 010 - 82028899

购书热线 010 - 64054588

免费咨询 800 - 810 - 0598

网 址 http://www.hep.edu.cn

http://www.hep.com.cn

经 销 新华书店北京发行所

印 刷 北京民族印刷厂

开 本 787×1092 1/16

印 张 33

字 数 810 000

版 次 2003 年 9 月第 1 版

印 次 2003 年 9 月第 1 次印刷

定 价 56.00 元

本书如有缺页、倒页、脱页等质量问题,请到所购图书销售部门联系调换。

前　言

本书主要是为了帮助生物学家理解序列与结构分析技术。我相信每位使用电脑程序的人都应该理解程序是如何运作的。因此，我的一个主要目的是让生物学家们体会程序背后所用的算法，这种算法是建立在何种假设上的，这些技术的局限以及使用这些技术的策略。为此，我尽可能使用简单的数值实例以避免复杂的公式。同时，我也希望，对于那些想更多了解与生物信息学相关的生物问题的计算生物学家，这本书能引起他们的兴趣。我希望这本书能成为一本实验参考书以及一本生物信息学课程的教科书，而不是成为一套序列分析程序的用户指南。

大多数章节中都包括一个流程图，用来介绍该章节中所讨论的各种技术的使用顺序。这类图表非常少见而且难以制作，需要假定和高度简化，以至于它们有可能并不适用所有的情况。我希望这些图表能对这些领域中的初学者有所帮助，而对于那些有经验的实践者，他们可以用自己的、也许更好的流程去达到同样的目的。

书中有很多相关网站和 FTP 站点可以应用这些技术或者下载程序。在某些情形中，比如常用的 BLAST 和 CLUSTALW 两个程序，我提供了大量的关于使用程序和分析结果的信息。然而，还有很多其他重要的序列和基因组分析的工具和方法，由于时间与篇幅所限，我只能尽可能多地介绍。我并未特别关注一些较为简单的序列分析，如寻找限制性位点、翻译序列以及组分分析，因为对这些工作，已经有大量的商业性和非商业性软件包。而商业性的基因组分析软件包才刚刚出现。

在写作本书时，我最大的感受就是，已发表的文献中所含的信息量之大已完全超过了我的概括能力。我试图全面涵盖序列及基因组分析中最重要的问题，但由于时间和篇幅的限制，仍有许多优秀的论文我没能引用。我对那些作出了有价值的贡献却未被提及的同行表示抱歉。由于印刷本的篇幅限制以及生物信息学所具有的不断变化的特点，那些书中没有包括的材料，所有书中引用过的网站的链接、例子和问题，都可以在与本书配套的网站中找到，网址是：http://www.bioinformaticsonline.org。

这个领域中的一个惯例引起了我特别的注意，那就是这个领域中的研究者，尤其是那些开拓者，心甘情愿地与他们的同行们分享成果。我认识几位早期研究者，他们使我获益匪浅，尤其是 David Lipman、Hugo Martinez（我曾经与他共度了一个学术休假年）和 Temple Smith，他们不仅取得了巨大的成就，更值得赞扬的是，他们率直无私地与同行分享他们的成果。正因如此，他们对序列分析领域在理论上以及商业上获得最终的成功功不可没。

本书的编写需要许多支持与帮助。本书一部分源于 1999 和 2000 学年度亚利桑那大学“生物信息学和基因组分析”课程的课堂笔记。很多学生提出了非常有用的建议并帮助纠错。其中，我特别感谢 Bryan Zeitler，他为本书作了很多修正。仍有遗漏的错误在本书配套的网站上已做校正。我要对以下各位表示感谢：Bill Pearson 为 FASTA 软件包提供资料；Julie Thompson 和

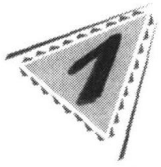
John Kececioglu 为第四章提供注解；Steve Henikoff 审阅第三章；Michael Zuker 在我撰写第五章时提供注解；Bill Montfort 为第九章的 PDB 文件提供资料；Roger Miesfeld 为第八章中的复杂基因调控提供实例；Jun Zhu 非常好地回答了我关于第三章中“贝叶斯区块排列”的问题。我所在的系对我的工作表现了极大的耐心与支持，允许我在长达三年的时间里缺席会议和专题讨论班，以便我完成其他章节的撰写与修改。在此期间，Rob Han 和 Juwon Kim 为我提供了大量文献和书籍，使我有更多的时间去消化那些信息。我的编辑——冷泉港实验室出版社的 Judy Cuddihy 在整个写作过程中一直在写作技巧方面指导我，十分耐心地给我一个通情达理的写作时间表，并鼓励我完成这本书。Elisabeth Cuddigy 核对了大多数的网站，仔细地检查了公式和例题中的运算并帮助完成了部分词汇表。我还要感谢出版社开发部的 Joan Ebert 和 Jan Argentine 以及制作部的 Pat Barker 和 Denise Weiss。

最后，我还要感谢我的妻子 Jennifer Hall。很多时候，本书的写作使我无暇顾及家事，她对此表示了忍耐和理解。

David W. Mount

郑重声明

目　录



历史简介与概述

序列分析方法的发展依靠众多具有不同科学背景的学者的贡献。本章对历史上许多重大进展进行简要地回顾，并对全书各章节作一概述。由于篇幅所限，很多贡献未能提及。本章参考文献中涉及许多早期及现在的参考书籍、论文、综述和杂志，可以为我们提供该领域更宽阔的视野。

最先收集的是蛋白质序列

蛋白质测序技术的发展(Sanger&Tuppy，1951)使得人们能对常见的蛋白质家族进行测序，例如对来自不同生物的细胞色素进行测序。Margret Dayhoff(1972，1978)和她在华盛顿特区的“美国生物医学研究基金会(National Biomedical Research Foundation，NBRF)”的合作者们，在20世纪60年代首先将这些序列数据库组合成一个蛋白质序列图谱集。他们的收集中心即为后来的“蛋白质信息资源(Protein Information Resource，PIR)”，曾被称为“蛋白质鉴定资源(Protein Identification Resource)”，(网址：http://watson.gmu.edu:8080/pirwww/index.html)。NBRF从1984年起开始负责维护该数据库。1988年，NBRF与慕尼黑蛋白质序列中心(Munich Center for Protein Sequences，MIPS)以及日本国际蛋白质信息数据库(Japan International Protein Information Database，JIPID)合作建立了一个PIR-国际蛋白质序列数据库(PIR International Protein Sequence Database)，(网址：http://www-nbrf.georgetown.edu/pir)。


Margaret Dayhoff

Dayhoff及其合作者根据蛋白质序列的相似程度，将蛋白质归为家族(family)和超家族(superfamily)。由此可以获得反映一组近缘蛋白质变化频率的表格。差异小于15%的蛋白质被挑选出来，以避免用氨基酸变化反映两条氨基酸序列而不是仅仅一条序列的变化时可能出现的偶然性。根据已排列的序列，可以构建一个系统树，以显示哪些序列亲缘关系最近并且共有同一支。一旦构建了这些树，就可以预测在不同生物中编码这些蛋白质的基因进化所产生的氨基酸变化(图1.1)。

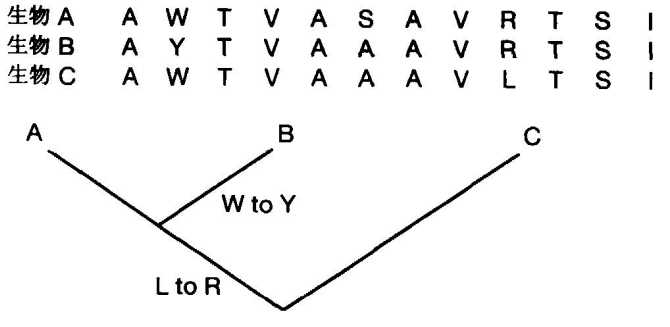
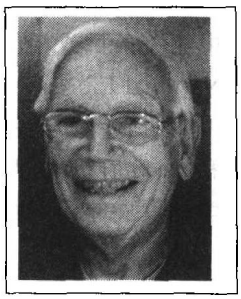


图1.1 相关蛋白质序列进化中的系统发育关系和可能的氨基酸变化的预测方法。3条高度保守的序列(A、B和C)来源于3种不同生物的相同蛋白质。这些序列非常相似，在进化中每个位置最多仅有一次变化。蛋白质差异为一到两个置换，因而可用于构树。一旦获得了这个树，就可以确定图中示出的两个氨基酸变化。图示的特定替换远比随机置换过程更有可能发生

随后，一个矩阵集（表格）——氨基酸突变被进化选择所接受的百分率或称 PAM 表——表明一个氨基酸变到所构系统树中另一个氨基酸的概率，用它可以显示两个序列的对应位置上哪个氨基酸最为保守。这些表格至今仍用于预测蛋白质间的相似性以及在数据库搜索中用于寻找与检索序列相匹配的序列。所用规则是：两个序列中的氨基酸愈一致、愈保守，则它们在进化过程中具有共同祖先基因的可能性愈大。如果这些序列非常相似，则蛋白质很可能有相同的生化功能及三维结构折叠。因此，Dayhoff 及其同事提出了几种用于现代生物序列分析的方法。他们提供了第一个蛋白质序列数据库以及用于蛋白质序列比较的 PAM 表格。氨基酸置换表格经常用于序列对位排列和数据库相似性检索，详见第三章和第七章。

DNA 序列数据库

DNA 序列数据库最早由 Walter Goad 及其在 GenBank 数据库和欧洲分子生物学实验室（EMBL，位于德国海德堡）的同事们在位于美国新墨西哥州的 Los Alamos 国家实验室（LANL）建立。已翻译的 DNA 序列也收录于美国生物医学研究基金会（NBRF）的 PIR 数据库。Goad 在 1979 年确立了 GenBank 原型系统；LANL 从 1982 年至 1992 年收集 GenBank 数据。GenBank 现在由美国国家生物技术信息中心（NCBI）（网址：http://www.ncbi.nlm.nih.gov）管理。EMBL 数据库于 1980 年建立（网址：http://www.ebi.ac.uk）。1984 年，日本 DNA 数据库（DDBJ）在三岛市建成（网址：http://www.ddbj.nig.ac.jp）。GenBank、EMBL 和 DDBJ 现在已组成国际核苷酸序列数据库合作体（网址：http://www.ncbi.nlm.nih.gov/collab），每日进行数据交换。PIR 也有类似的合作。



Walter Goad 许多类型的序列数据库见每年第一期 Nucleic Acids Research Genbank 中序列的增长数目见 http://www.ncbi.nlm.nih.gov/GenBank/geneba

最初，一个序列条目仅由计算机文件名和 DNA 或蛋白质序列文件构成。之后，扩充了序列的功能、突变、编码蛋白、调控位点及文献等信息。这些信息都以数据库的格式与序列放置在一起，以便用户查询各种类型的数据。对多种数据库及其格式的讨论见第二章。

GenBank 和 EMBL 核苷酸序列数据库中的条目数随每日更新而持续地快速增长。注释所有这些新序列是一个费时、艰苦并且可能出现错误的过程。随着时间的推移，这一过程逐渐自动化，但带来精确性和可靠性的问题。1997 年 12 月，GenBank 中共有 1.26×10^9 个碱基，1999 年 4 月达到了 2.57×10^9 个碱基，2000 年 9 月已达到 1.0×10^{10} 个碱基。尽管贮存序列数呈指数增长，但一些高效的搜索技术可帮助人们调用这些序列。

为了减少数据库搜索中的匹配数，仅列出一致序列中一个代表序列的无冗余数据库（nonredundant database）已经出现。然而，大多数序列数据库里仍包含大量相同基因或蛋白质的条目，其内容可能源于序列片段、专利、来自不同数据库的备份以及其他类似序列。

从公共数据库中调用序列

通过网页可以检索主要的序列数据库（GenBank 和 EMBL 等），这是提供序列数据库访问的


David Lipman

重要步骤。在NCBI，这种技术的早期实例是由D. Benson和D. Lipman及其同事所开发的一个称为GENINFO的菜单驱动程序。该程序可以借助索引序列数据库快速寻找生物学家查询的条目。随后，NCBI开发了ENTREZ(网址：http://www.ncbi.nlm.nih.gov/Entrez)，这是一个由窗口界面最终发展为网页界面的程序。这些程序的设计思想是为序列数据库提供易操作的界面和灵活的搜索。

在这些主要的数据库中，序列条目包含一些有关该序列的附加信息，如序列编号或索引号、序列名和别名、相关基因的名称、调控序列的类型、序列来源于何种生物、文献以及已知的突变。ENTREZ能访问所有这些信息，从而允许对整个序列数据库进行快速搜索，找出符合一个或多个条件的序列(在ENTREZ中被称为近邻序列)。当需要在一个数据库中进行单一或多项检索时，简单模式搜索程序只能找到与查询条件完全匹配的结果。相比之下，ENTREZ可以对相似、相关条件或者包含了多个选项的复合条件进行搜索，并将搜索到的项目按与检索条件匹配的可能性的大小顺序列出。ENTREZ原先已允许直接访问DNA和蛋白质的序列数据库及文献，甚至可以得到不同或同一数据库中相关条目或相似序列的索引。最近，ENTREZ提供了对整个MEDLINE，即美国医学图书馆(National Library of Medicine，NLM，位于华盛顿特区)的完整书目数据库的访问权，以及对其他一些数据库(如生物系统发育数据库和一个蛋白质结构数据库)的访问权。访问对任何用户——私人、政府机构、产业实体和研究机构都是免费的，这是全体NCBI职员的决定，它对整个生物医学的发展起到不可忽视的促进作用。目前，NCBI每天要处理数百万个独立的访问。

数据库查询程序(如ENTREZ)为我们获取不断增加的序列和生物医学杂志提供了便利。然而，无论使用何种自动技术，我们必须清楚，数据库搜索也许并不能获得所有的相关资料，一些重要的条目也许会错过。每个数据库的登录在某些情形下总是需要手工操作的，这就难免增加拼写错误和其他问题出现的频率。有时，一篇应该在数据库中出现的参考文献可能没有找到，这可能是因为搜索条件中出现了拼写错误、条目的确不存在或者其他更为复杂的原因。如果穷尽搜索和试探方法均告失败，请将问题报告给程序管理员或者系统管理员。

序列分析程序

由于DNA测序是检测一块测序凝胶上的一组峰值(A、G、C、T)，因而这一过程很容易出错，这取决于数据质量。

DNA测序方法是由Maxam和Gilbert(1997)和Sanger等(1997)发展起来的。第二章将作详细介绍。

在20世纪70年代末，越来越多的DNA序列被测定，人们不断开发计算机程序，用多种途径来分析这些序列。1982年和1984年，Nucleic Acids Research出版了两期特刊，专门介绍计算机在序列分析中的应用，包括将大型计算机程序移植到当时最新的微型机上。稍后，威斯康星大学J. Devereux领导的遗传学计算机组(Genetics Computer Group，GCG)开始提供在VAX型计算机上使用的序列分析程序。GCG最终走向了商业化(http://www.gcg.com/)。其他提供微机序列分析程序的公司(如In-

telligenetics 和 DNAStar 等)大约也在这一时期出现。一些实验室也开发免费或低价共享的计算机程序。例如,为了简化数据采集,华盛顿大学的 Phil Green 和他的合作者们开发了 PHRED 和 PHRAP 程序(Ewing 和 Green,1998;Ewing 等,1998),以辅助序列数据阅读和处理。PHRED 和 PHRAP 现在由 CodonCode 公司发行(网址:http://www.codoncode.com)。

这些商业化和非商业化程序至今仍被广泛使用。此外,很多网站现在也可以进行多种序列分析。它们对学术研究机构是免费的,对商业用户适度收费。下面简要综述序列分析方法的发展。

序列比较的点阵或图形法

1970 年,Gibbs 和 McIntyre(1970)提出了一种比较两个氨基酸或核苷酸序列的新方法。该方法作一个图,一条序列横排在上首,另一条纵列在左端。两个序列在任何位置上若出现相同值,就在两个序列对应位置的交叉位置上标注一个点(图 1.2)。在结果图上,排列成对角线的点列体现出两条序列的相似性,或两个序列间相同的字符串。只要使用较小的点,长序列也能画在一张纸上。

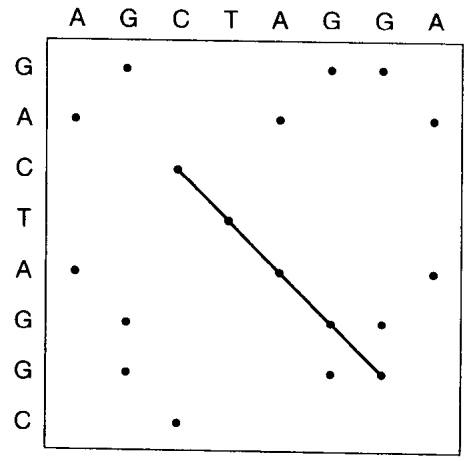


图 1.2 两个 DNA 序列比较的一个简单点阵(两个序列分别为 AGCTAGGA 和 GACTAGGC)。这些点的对角线模式显示出两个序列间有一段相似的序列 CTAGG

点阵方法可以快捷地找到序列间存在的插入或缺失,因为它们的存在使对角线在水平或垂直方向上发生了位移。对一个序列进行自比较,可以找到序列中存在的重复序列。这种重复序列可以是同向、反向或者回文的。序列自比较方法可以揭示一些特征,如染色体间的相似性、串联基因、蛋白质序列中的重复功能域、序列复杂度低而易发生重复的区域以及 RNA 中的自补序列。当相似性非常低时,有可能图中无法出现对角线。Gibbs 和 McIntyre 计算了所有可能的对角线,并将它们与随机序列间得出的结果进行比较,用以找出最显著的对位排列。

后来,Maizel 和 Lenk(1981)发展了各种过滤和彩色显示模式,大大增加了点阵方法的有效性。序列比较的点阵表示法仍然在 DNA 和蛋白质序列相似性分析以及基因和极长染色体序列中的重复分析中发挥重要的作用。第三章中将作详细介绍。

用动态规划法进行序列对位排列

尽管点阵方法可以用来检测序列的相似性,但当序列相似性被某些区域打断(如插入或缺失)时,该方法有可能会失效。因此,人们希望能够改造点阵方法,以提供两个序列间最可能的排列,称为最优排列。这样的排列是将序列按行写下来,把序列间匹配的项写在同一列里,那些无

法匹配的项可以作为错配，也写在同一列里，或者加上一个间隔作为插入（或者看作为另一个序列的缺失）（图 1.3）。为了寻找最优排列，所有可能的匹配、插入、缺失都要被考虑。这样做的运算量非常之大，如一个长度为 300 的蛋白质序列需要 10^{88} 次比较（Waterman，1988）。

序列 A：A G △ △ C D E V I G
序列 B：A G E Y C D △ I I G

图 1.3 两个序列的对位排列。从中可以看到匹配、错配和间隔（△）

为了简化这一工作，Needleman 和 Wunsch（1970）将问题转化为用渐进的方式构造一对氨基酸序列的对位排列。这种方法从一对序列的末端开始，每次前进一个氨基酸对，允许匹配对、错配对和额外的氨基酸（插入或缺失）以不同方式组合在一个序列中。这种方法在计算机科学中被称为动态规划。Needleman-Wunsch 方法产生两套结果：（1）每一个可能的对位排列。每种排列都包括了所有的匹配、错配和每一个插入（或缺失）的所有组合。（2）一个用于评判对位排列优劣的记分系统。可以根据记分的高低来决定何种排列方式是最好的。每一个匹配对得 1 分，错配对得 0 分，一个间隔扣 1 分。对这些数字求和，可以获得这个排列的总分。得分最高的排列被定义为最优排列。

产生所有可能排列的过程是这样的：在与点阵图非常相似的矩阵中，从其中一个序列的末端相应位置开始，在所有匹配的位置上持续移动（图 1.4）；在矩阵的每一个位置上，考虑任一序列中所有可能的起始位置以及匹配、错配、插入和缺失的所有组合，最高得分就记在这个位置上；找出图中得分最高的位置，然后从这一位置在图中沿着产生这个高分的所有位置的路径反推，就能得到最优排列方式。在已排列的两个序列中，对应于矩阵中那条路径的点都是匹配的。

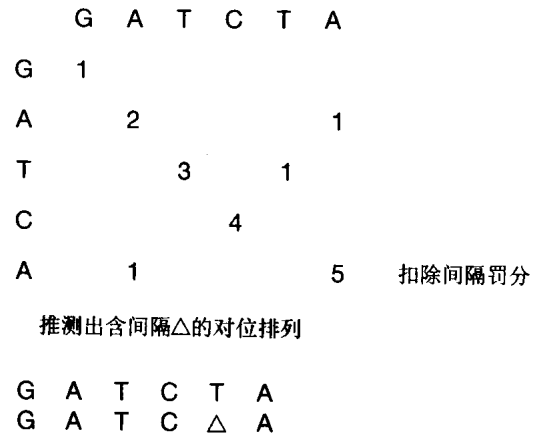
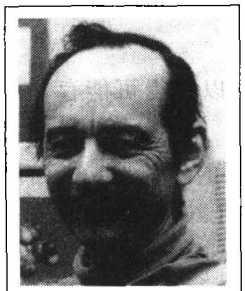


图 1.4 一个简化的 Needleman-Wunsch 对位排列例子。两个序列分别为 GATCTA 和 GATCA。首先，给两个序列中的所有匹配对分别记 1 分，错配对记 0 分（没有画出来）。然后，将同一对角线上的每一个“1”求和，于是得到了一个总分“4”。目前，本行无法继续延长而达到 5 分，但我们可以在 GATCA 中加入一个间隔，变为 GATC△A，△就是间隔的位置。这时，我们就可以在右下角的位置得到匹配，获得最高分 5 分了。但加上这个间隔后，我们还必须从总分中扣除一个罚分。从得到最高分的位置开始，通过对所有为这一分值作出贡献的位置反推，即可得到最优排列

在序列间寻找局部对位排列

上述方法都是在两个序列的整体范围内寻找最优对位排列，这类排列称为全局对位排列。Smith 和 Waterman（1981a，b）认为，DNA 和蛋白质序列中大多数具有重要生物学意义的区域都是排列很好的区域，而剩下的那些没有亲缘关系，也不能很好排列的区域大多不那么重要。因此，他们对 Needleman-Wunsch 算法作了一个重要修改，称为局部对位排列或 Smith-Wa-


Mike Waterman


Temple Smith

terman(或 Waterman-Smith)算法，来确定这些局部区域。他们还发现，任何规模的插入或缺失可认为是序列在进化过程中发生的改变，因此可以调整排列方法来适应这种改变。最终，他们用数学方法证实了动态规划能够在两个序列间找到一个最优的排列。这一算法的细节详见第三章。

两种互补测度可以用于对两序列排列进行记分，得到一个相似性分值和一个距离分值。如图 1.3 所示，在排列的每一列里有 3 种情况，分别是匹配、错配和间隔。例如，在一个简单的记分系统中，任何一种匹配都得 1 分，将已排列的序列中的匹配值相加，得到相似性分值。这个分值要除以匹配对和错配对的总数(间隔往往被忽略)。这一相似性记分系统是生物学家最为熟悉的，由 Needleman&Wunsch 提出并由 Smith&Waterman 开始运用。另一种记分系统用来计算两个序列间的距离分值。对一个序列变为另一个序列所需的替换数求和，即可获得距离分值。这个记分对估计基因或蛋白质间的进化距离进而用于推断系统发育(进化)是最有用的。这一方法主要依靠数学家们(尤其是 P. Sellers)的工作。距离记分通常是将所有的匹配数除以匹配数与错配数之和。运算结果表示出一个序列变为与它比较的另一个序列需要的替换数(不考虑间隔)。因此，在图 1.3 所示的例子中，共有 6 个匹配和 1 个错配。如果给定匹配分值为 1，则排列的相似性记分为 $6/7=0.86$，而距离记分为 $1/7=0.14$。用这样一个简单记分方案，相似性与距离记分之和为 1。注意，序列长度之和等于两倍的匹配数与错配数及插入(或缺失)之和。因而，在本例中，序列长度之和是 $8+9=2\times(6+1)+3=17$。通常，更为复杂的记分系统可用于完成具有特定含义的对位排列，并用概然率或概率记分来评价不同的排列(详见第三章)。不过，相似性与距离记分的互补关系仍然存在。

序列对位排列所面临的一个困难是：如何决定某个排列是否具有显著意义。这个排列方式的记分是否揭示了两个序列间的相似性？或者说，这个记分能否从两个没什么亲缘关系的序列(例如其中一条为计算机产生的随机序列)中得出？这个问题由 Karlin 和 Altschul(1990，1993)提出，在第三章中将作详细介绍。

对远缘或随机序列的记分分析表明，这种情况下的得分有时会比预期的正态分布中的分数要高很多。相反，具有偏尾分布的记分已知为各种分布下记分的极值。这种分析提供了一种概率评估方法，针对两个近缘序列间得到的记分可能与两个与它们长度相同的远缘或随机序列的排列记分相同的情况。这对评价检索序列与数据库中序列间的匹配程度具有特殊意义，第七章中将作进一步的讨论。在这种情况下，一个排列的记分必须将数据库中序列比较的次数考虑在内。因此，如果一个蛋白质检索序列与数据库中的蛋白质序列进行比较获得一个记分，而与一个远缘序列比较所产生同一记分的概率为 10^{-7}，若共有 80 000 个序列进行比较，则最高期望记分为 $10^{-7}\times8\times10^{4}=8\times10^{-3}=0.008$。期望值在 0.02～0.05 之间时视为是显著的。即使我们得到这一期望值，仍需对排列仔细检查，因为太短的排列、不真实的氨基酸匹配以及重复氨基酸序列都会导致排列的置信度降低。

多重序列对位排列

除了两个序列的对位排列外，人们还发展了 3 个及 3 个以上的序列排列方法(早期的实例见 Johnson 和 Doolittle，1986)。这些方法高度依赖于计算机技术，而且通常以两序列排列为基础。常用的软件有 GCG 的 PILEUP 程序(http://www.gcg.com/)和 CLUSTALW(Thompson 等，1994)(网址：http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html)。一旦一组相关的分子序列(一个家族)间的排列完成，就可以鉴定出高度保守的区域(Gribskov 等，1987)。这些保守区域为家族共有的，可用于鉴定家族中的其他成员。在多重序列排列中，称为 PROFILE 和 POSITION-SPECIFIC SCORING MATRIX(PSSM)的两个矩阵是非常重要的计算工具。

多重序列排列也可作为进化建模的起点。检验已排列序列的每一列，可以获得可能存在的系统发育关系或者再现所观测变化的系统树。

另一种多重序列排列的方式是无需首先排列序列，而是寻找所有待排列的 DNA 或蛋白质序列的共有模式(Stormo 等，1982；Stormo 和 Hantzell，1989；Staden，1984，1989；Lawrence 和 Reilly，1990)。对蛋白质序列，这些模式可以定义为一个保守结构或功能域。对 DNA 序列，这些模式则可能定义为一个在启动子区域的调控蛋白结合位点或者是 RNA 分子中的一个处理信号。统计学与非统计学方法均广泛应用于寻找共有模式。这些方法对序列进行分拣，试图在每条序列中找出一系列邻接字符，以便为排列提供最高的匹配数目。神经网络、隐马尔可夫模型以及期望极大化与吉布斯取样方法都是常用的方法(Stormo 等，1982；Lawrence 等，1993；Krogh 等，1994；Eddy 等，1995)。第四章中将介绍这些方法并给出实例。

RNA 二级结构预测

除了预测蛋白质结构的方法外，计算机预测 RNA 二级结构的方法在很久以前也已开始发展。如果一个 RNA 分子的序列与相反化学方向上的序列存在回文互补，则可形成一个“发夹”结构的区域(图 1.5)。

Tinoco 等(1971)用小寡核苷酸分子合成了这些对称区域。他们使用一个能值表，对这个模型中堆叠碱基对联合自由能以及环状结构的失稳作用进行了估算，以推测这些对称区域的稳定性(Tinco 等，1971；Salser，1978)。单链环状结构和其他未配对的区域减少了预测的能值。随后，Nussinov 和 Jacobson(1980)设计了一种快速预测 RNA 分子最高配对碱基数的计算机方法。这种方法与用于序列排列的方法使用的是同一种动态规划算法。Zuker 和 Stiegler (1981)对这一方法进行了改进。他们增加了分子约束和热力学信息，以预测在能量上最稳定的结构。

RNA 的结构建模的另一个重要用途是用于构建 RNA 分子数据库。Woese(1987)实验室所建的核糖体 RNA 数据库是最为重要的一个(网址：http://www.cme.msu.edu/RDPhtml/index.html)。RNA 二级结构预测将在第五章中讨论。基于这些 RNA 序列，人们可以进行对位排列、结构建模以及系统发育分析，从而探索生物的进化关系。

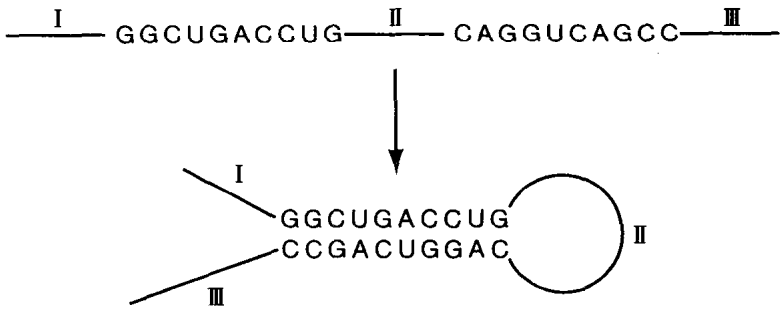


图 1.5 单链 RNA 分子折叠形成一个"发夹"二级结构。图中显示序列的互补部分。它们可以进行碱基配对以形成双链结构。G/C 之间的配对由于存在 3 个氢键而具有最高能量；A/U 和 G/U 配对因分别有 2 个和 1 个氢键而产生能量递减

应用序列揭示进化关系

同一个核酸或蛋白质序列家族中的不同成员可以为进化生物学研究提供极为宝贵的信息。随着序列信息的不断丰富，我们可以利用序列信息来寻找一些古老基因的踪迹（如核糖体 RNA 和一些蛋白质）；从推测进化树到利用序列来发现新物种（Barns 等，1996）。不同的进化历史导致了基因多样性，反映出种间遗传物质的转移。其他类型的系统发育分析可以用来鉴定同一家族中的基因，这些基因由进化后裔相关联，称为直系同源。基因复制事件可以形成一个基因的两个拷贝，称为并系同源。很多复制事件可以产生一个基因家族，每个成员在功能上变化都很小。当然，也可能出现全新的功能。一旦产生了对位排列并获得排列记分后，就可以找到最接近的序列对，它们有可能出现在进化树外部分支上，如图 1.1 中的 A 和 B 所示。图中的 C 序列是下一个最相似的序列，用树的下一个分支表示。继续这一过程，就可以对某个基因的进化型式进行预测。一旦构建一个树后，可以推断出那些发生在树枝上的序列变化。

序列对位排列是构建系统树的起点。对每对序列，相似性记分指示出哪些序列在亲缘关系上最为接近。然后，根据这些记分就能推断出计测序列间变化（距离）的最佳树（Fitch&Margoliash，1987）。最常用的构树方法是邻接法（Saitou&Nei，1987），第六章中将会详细介绍。其他可供选择的方法是，若一个可靠的多重序列排列已经完成，也可构建与已排列序列每列观测到的变化最一致的树。这样获得的树意味着因变化而赋予的罚分最少（最大简约树）（Felsenstein，1988）。

在进行系统发育预测时，我们必须考虑几种树具有相同结果的概率。因此，可以用显著性检验来确定序列对某个树的支持程度（Felsenstein，1988），这些也将在第六章中讨论。

相似序列数据库检索的意义

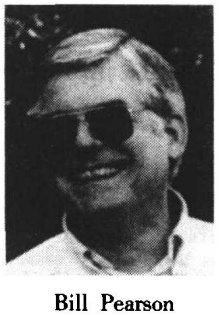
随着 DNA 测序成为一种常见的实验室工作，人们可以对一些具有重要功能的基因进行测

序，以了解基因产物的生化特性。以在动物中引起癌症的逆转录病毒编码的 *v-sis* 和 *v-src* 致癌基因为例，通过比较预测的病毒产物的序列与所有已知的蛋白质序列，Doolittle 等(1983)以及 Barker&Dayhoff(1982)得到了一个令人震惊的发现：这些基因都起源于细胞基因。*sis* 蛋白质有一段序列与哺乳动物的血小板生长因子(PDGF)非常相似，而 *src* 则与哺乳动物中的环腺苷酸依赖的激酶中的催化链相似。因而，可以推测逆转录病毒可能通过某种基因交换途径从寄主细胞中获得并在病毒感染其他动物时产生一个蛋白突变型，部分替代普通蛋白的功能。此后，分子生物学家们分析了愈来愈多的基因序列，并根据序列的相似性发现了很多生物共有的相似基因。

从大肠杆菌(*Escherichia coli*)和酿酒酵母(*Saccharomyces cerevisiae*)等模式生物中获得的遗传和生化信息，使这类检索变得更为便利。在这些模式生物中，广泛的遗传分析已经揭示了基因的功能，这些基因的序列也已测定。如果在别的生物(如一种农作物)中发现的新基因与模式生物(如酿酒酵母)中的基因具有相似的序列，可以预测这个新基因可能与模式生物中的基因具有相同的功能。这类检索工作目前已经非常普遍。许多计算机程序——如 FASTA (Pearson&lipman,1988)和 BLAST(Altschul 等,1990)等——对检索工作帮助极大。

在 BLAST 程序以及其他序列相似性检索中运用的方法将会在下一段和第七章进一步介绍。

数据库检索中的 FASTA 和 BLAST 方法


Bill Pearson

随着实验室中获得的序列数目增加，对新序列与数据库中序列逐一比较的计算机程序的需求也日益增长。对病毒致癌基因功能的鉴定就是一个成功实例。Needleman-Wunsch 的动态规划方法过去因计算机运算速度太慢而无法应用。今天，随着计算机运算速度的大幅提升，该方法可被实际运用。Pearson&Lipman (1988)开发了 FASTA 程序，可以在足够短的时间内对数据库进行相似性检索。FASTA 可以快速地从数据库中任意序列间找出与新序列相似的短串。每个序列首先被分解成只有若干字符长度的串，这些串被编入一张表中以标明它们在序列中的位置。如果某个或几个串在两个序列中同时出现，尤其是当几个同时出现的串彼此相接时，就意味着两个序列在这些区域相似。Pearson (1990,1996) 改进了 FASTA 方法。

Altschul 等(1990)编写了一个更快速的序列数据库检索程序，称为 BLAST。这一方法在国家生物技术信息中心(NCBI)的网站上得到广泛的应用(网址：http://www.ncbi.nlm.nih.gov/BLAST)。BLAST 服务器也许是世界上应用最广泛的序列分析工具，它能提供几乎所有序列的相似性检索。同 FASTA 一样，BLAST 也是首先制作一张每个序列中短序列串的表格，不同的是它还决定了哪些串更重要，能更好地反映两个序列间的相似性。然后将检索限于这几个关键字段及相关字段(图 1.6)。现在，有同时用于核酸和蛋白质数据库的 BLAST 版本，可以先将 DNA 序列翻译成蛋白质序列，随后与蛋白质序列数据库中的序列进行比较(Altschul 等,1997)。BLAST 的最新改进包括两个方面：一是 GAPPED-BLAST，它将原始 BLAST 版本的检索速度提高了 3 倍，也使检索的匹配结果得到了增加；二是 PSI-BLAST，它可以通过在剩余序列中反复检索与查询序列和先期已获得的序列相匹配的序列，从而发现更多与待检索蛋白质序列距离较